



TRUY VẤN ẢNH SỬ DỤNG CÂU TRUY VẤN KẾT HỢP ẢNH VÀ CÂU MÔ TẢ TĂNG CƯỜNG TIẾNG VIỆT

Sinh viên thực hiện
Nguyễn Trường Phát - 17520880

Giảng viên hướng dẫn
TS. Nguyễn Vinh Tiệp



KẾT HỢP ẢNH VÀ CÂU MÔ TẢ TĂNG CƯỜNG TIẾNG VIỆT CHO TRUY VẤN ẢNH

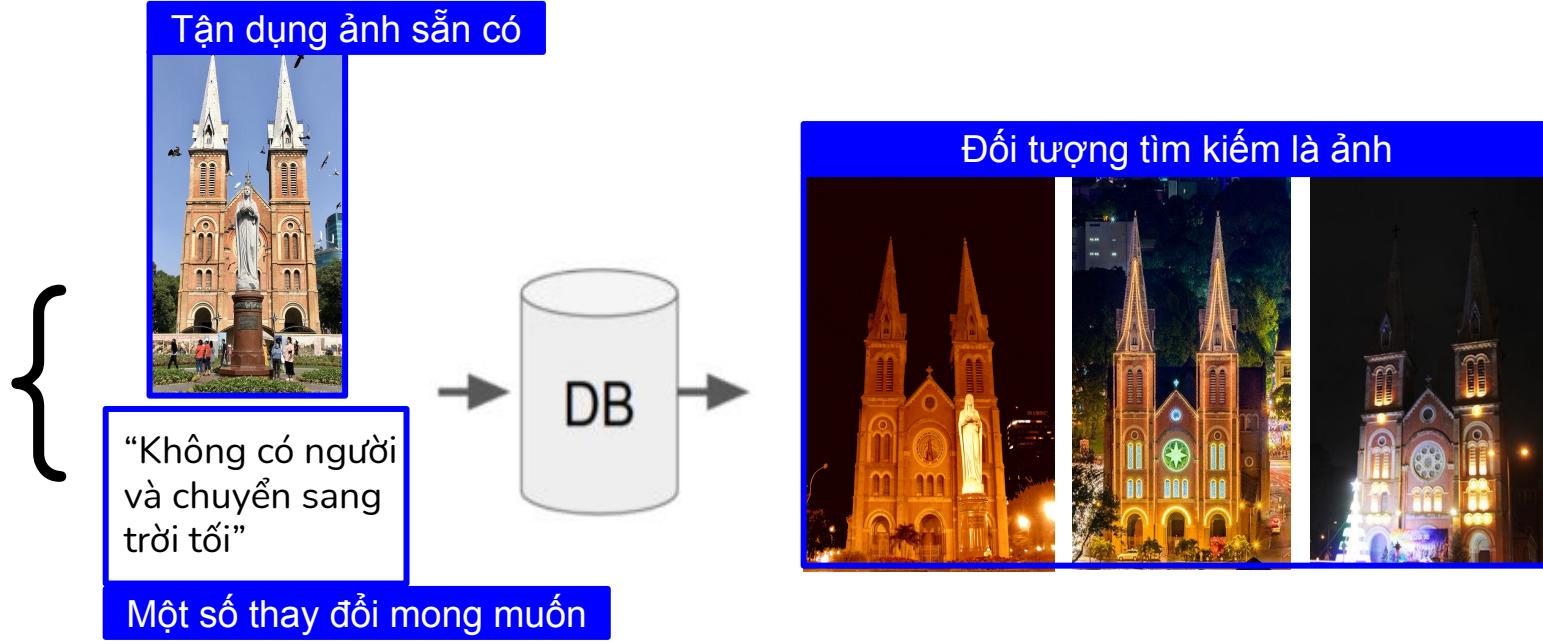
Sinh viên thực hiện
Nguyễn Trường Phát - 17520880

Giảng viên hướng dẫn
TS. Nguyễn Vinh Tiệp

Thành phố Hồ Chí Minh, tháng 2, năm 2021

Nội dung chính

1. Giới thiệu bài toán
2. Các hướng nghiên cứu liên quan
3. Hướng tiếp cận
4. Kết quả thực nghiệm
5. Kết luận và hướng cải tiến



→ Đây là một giao thức truy vấn có tiềm năng sử dụng cao do tính thuận tiện, tuy nhiên chưa được làm cho tiếng Việt

Những đặc thù ở ngôn ngữ tiếng Việt

1. Tiếng Việt phù hợp với các mô hình Học máy do sở hữu lượng từ phân loại rất phong phú

VD:

- “cái” - chỉ đồ vật Cái bàn - Table
 - “quyển” - chỉ đồ vật giống sách Quyển sổ - Notebook
 - “con” - chỉ con vật Con mèo - Cat

→ Cho thêm thông tin về danh từ đứng sau, tiềm năng ứng dụng Tiếng Việt trên các mô hình Học máy

Đinh-Hòa Nguyễn, "Tiếng Việt không son phấn", John Benjamins Publishing Company, 1997.

Những đặc thù ở ngôn ngữ tiếng Việt

2. **Tuy vậy**, Tiếng Việt tồn tại sự **nhập nhằng khoảng trắng**

Tiếng Anh (đa số) sử dụng khoảng trắng để phân cách các từ với nhau còn tiếng Việt thì chỉ để phân cách các âm tiết với nhau

VD: máy tính, thức ăn,..

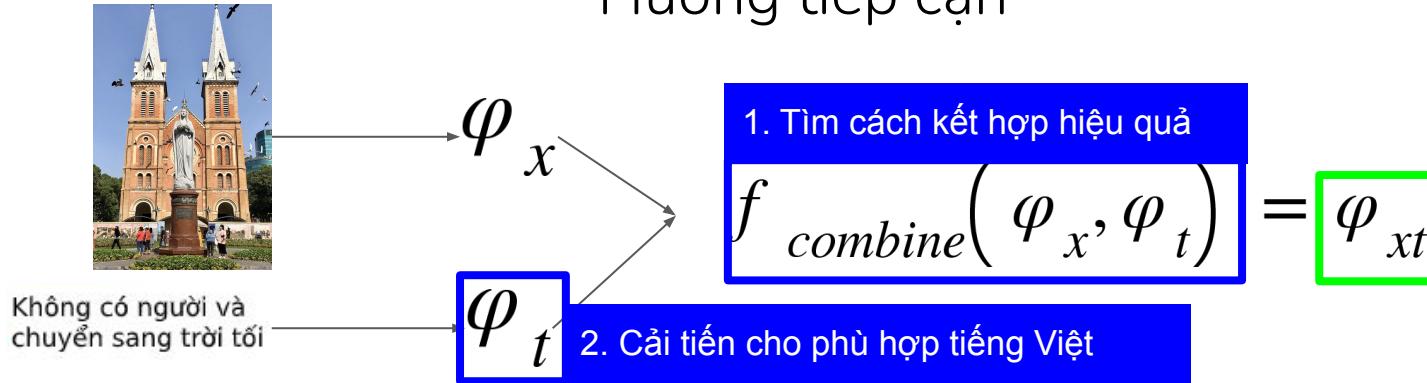
→ Động lực để **thiết kế lại phương pháp** cho phù hợp với **đặc thù** trên của tiếng Việt

Đinh-Hòa Nguyễn, "Tiếng Việt không son phấn", John Benjamins Publishing Company, 1997.

Những đóng góp chính

- 1 Báo cáo phương pháp **Text-Image Residual Gating** cho bài toán truy vấn ảnh dựa trên ảnh và câu mô tả
- 2 Cải tiến phương pháp **Text-Image Residual Gating** cho phù hợp câu mô tả là tiếng Việt
- 3 Xây dựng tập dữ liệu tiếng Việt **CSS-VN** để phục vụ quá trình nghiên cứu
- 4 Xây dựng **công cụ dịch** dựa trên **cây cú pháp**

Hướng tiếp cận



Các hướng nghiên cứu liên quan

FiLM

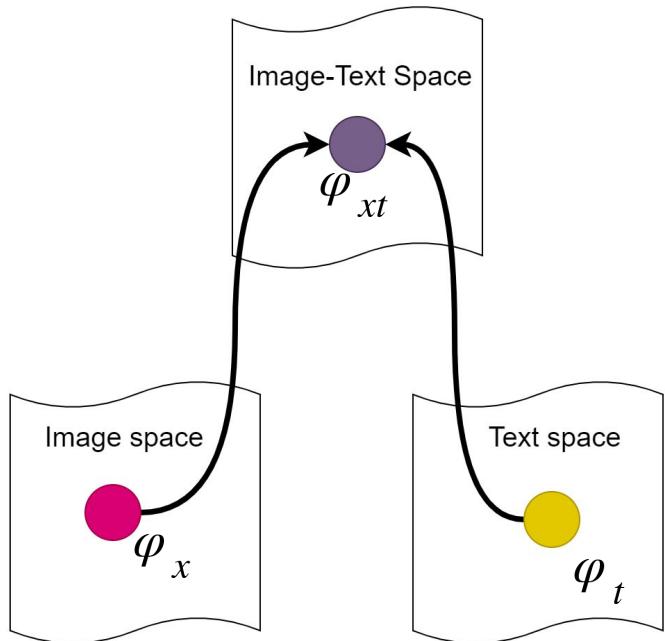
Parameter
Hashing

Concatenation

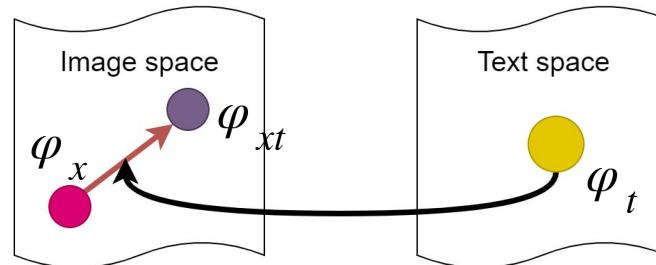
1. Chỉ áp dụng trên bài toán **Hỏi đáp trên ảnh** (*Visual Question Answering*)
2. Sử dụng các phép biến đổi đơn giản, **không gian biểu diễn bị hạn chế**

Vo, Nam, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. "Composing text and image for image retrieval—an empirical odyssey." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Các phương pháp kết hợp khác



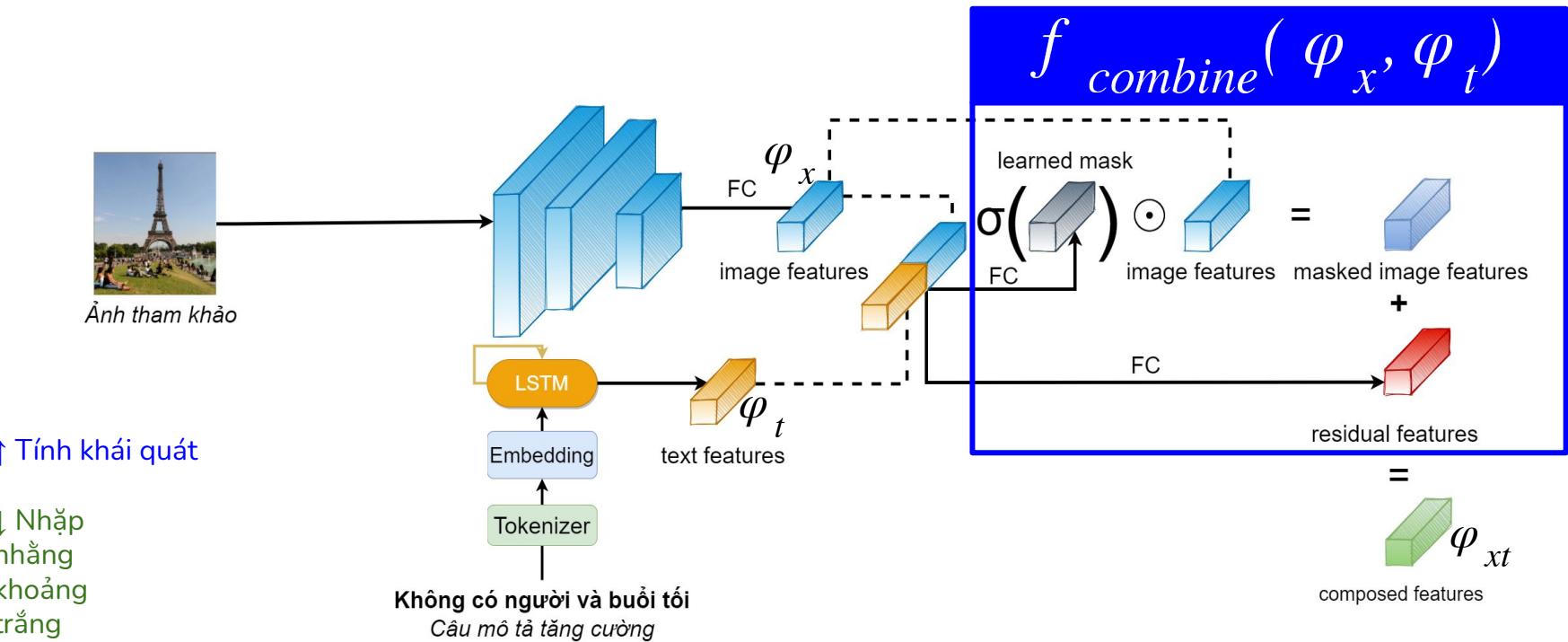
Phương pháp TIRG



→ Cách biểu diễn **phù hợp hơn** cho truy vấn ảnh

Vo, Nam, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. "Composing text and image for image retrieval—an empirical odyssey." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

Text-Image Residual Gating



Vo, Nam, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. "Composing text and image for image retrieval—an empirical odyssey." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

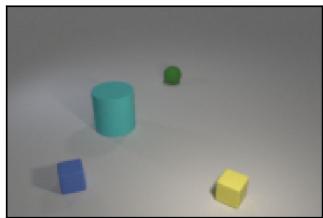
4. Kết quả thực nghiệm

Tập dữ liệu CSS-VN

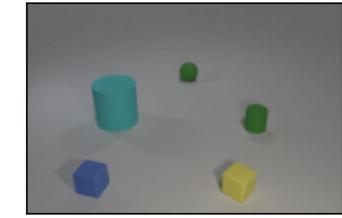
Từ phân loại (classifiers)

Gồm:

- **3 thao tác**: thêm, biến, xóa.
- **4 loại đối tượng**: khối, khối lập phương, khối cầu, khối trụ.
- **8 màu**: màu nâu, object ; cube sphere cylinder
- **9 vị trí**: bên trái, bên phải, trung tâm, góc phải phía dưới,..
- **2 kích thước**: lớn, nhỏ

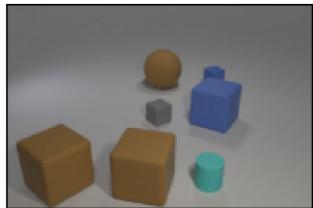


“thêm khối trụ màu xanh lá vào bên phải”



Một số mẫu khác trong tập dữ liệu CSS-VN

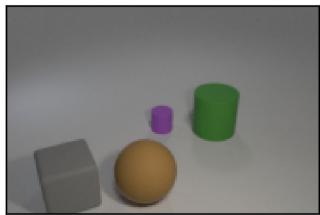
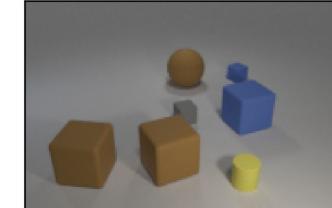
Ảnh tham khảo



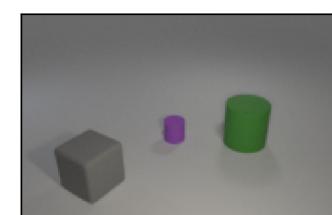
Câu mô tả

“biến khối trụ góc phải dưới thành màu vàng”

Ảnh mục tiêu



“xóa khối cầu màu nâu chính giữa phía dưới”



Độ đo đánh giá

Một mô hình truy vấn gọi là tốt khi mô hình đó trả về **tất cả kết quả phù hợp** ở các vị trí **đầu tiên**

$$R@k = \frac{1}{|U|} \sum_{u \in U} \frac{|Rel_u @ k|}{|Rel_u|}$$

Recall at K: tỉ lệ tất cả kết quả phù hợp được trả về ở K vị trí đầu tiên
Chúng tôi sử dụng R@K ở 5 mức xếp hạng: 1, 5, 10, 50 và 100

Kết quả xây dựng mô hình TIRG-PhoBERT cho CSS-VN

Cấu hình	R@1	R@5	R@10	R@50	R@100
CSS trên TIRG (công bố)	73.1	KCB*	KCB	KCB	KCB
CSS trên TIRG (tái hiện)	71.11	90.62	94.13	98.19	99.09
CSS-VN trên TIRG	78.03	94.91	97.18	98.28	99.57
CSS-VN trên TIRG-PhoBERT	78.05	96.30	97.91	99.44	99.66

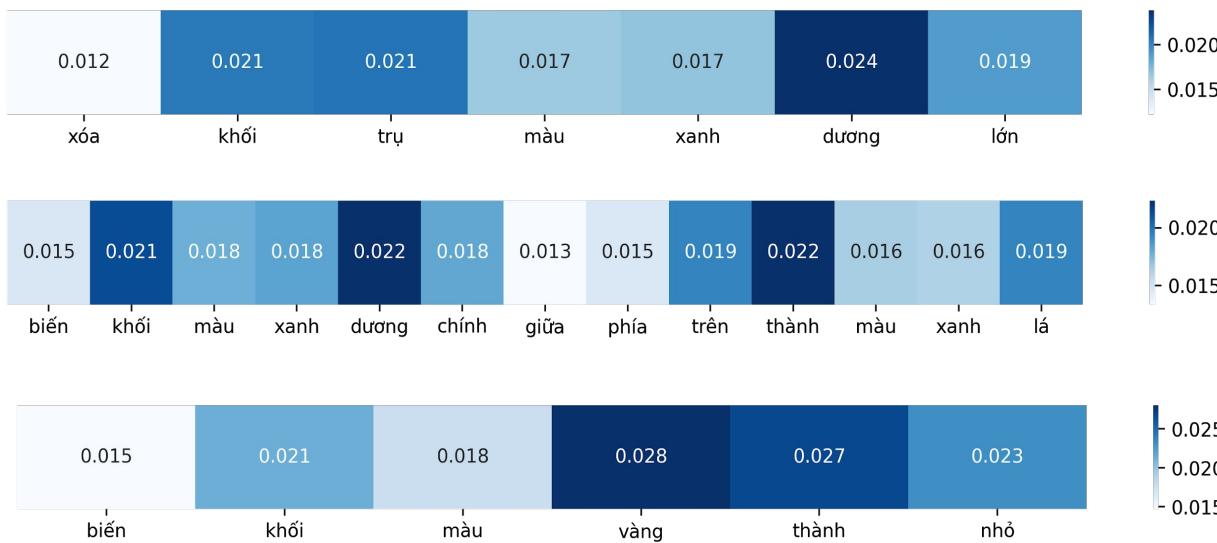
*KCB: Không công bố

R@1 trên TIRG-PhoBERT của các đối tượng ở CSS và CSS-VN

CSS	CSS-VN
“object” 81.03%	“khối” 82.64%
“cube” 53.99%	“khối lập phương” 65.13%
“sphere” 80.69%	“khối cầu” 83.44%
“cylinder” 57.42%	“khối trụ” 73.03%

→ Nhờ sự xuất hiện của classifier “khối” đã cho thêm **thông tin về từ loại**, cải thiện kết quả mô hình hóa câu

Trực quan hóa LSTM



→ LSTM đã mô hình hóa tốt các câu mô tả tiếng Việt, những từ quan trọng trả lời cho các câu hỏi “Khối gì?”, “Vị trí nào?”, “Kích thước gì?” và “Màu gì?”

Đánh giá tính thích ứng bằng CSS-VN-augmented

Từ gốc

“nhỏ”

Từ thay thế

“bé”

Ví dụ: Xóa khối cầu **nhỏ** **bé**

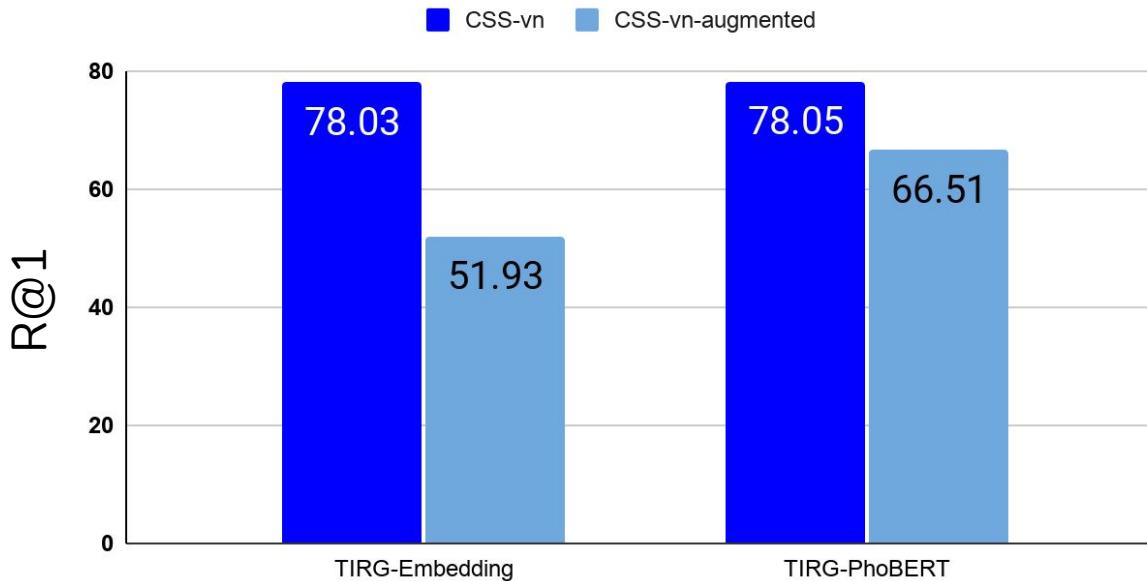
“lớn”

“to”

Ví dụ: Biến khối cầu **lớn** **to** thành màu đỏ

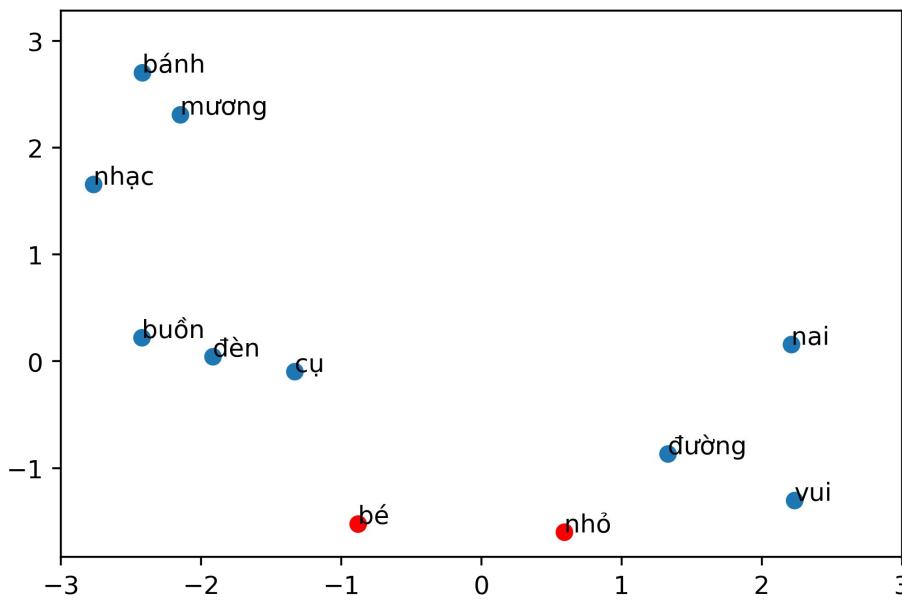
→ Để đánh giá tính thích ứng mô hình **TIRG-PhoBERT** với từ mới, ta tạo một tập dữ liệu mới với một số thay đổi nhỏ.

So sánh TIRG-Embedding và TIRG-PhoBERT



→ Mô hình với lớp **Embedding giảm R@1** trầm trọng khi đánh giá trên tập **CSS-vn-augmented**, còn **TIRG-PhoBERT** thì **robust** hơn.

Biểu diễn từ trong không gian PhoBERT



→ Các **từ tương đồng** sẽ được ánh xạ về các **biểu diễn tương đồng nhau** trong không gian PhoBERT

Tâm quan trọng của bộ tách từ RDRSegmenter

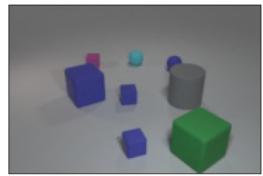
Cấu hình	R@1	R@5	R@10	R@50	R@100
Sử dụng RDRSegmenter	78.03	94.91	97.18	98.28	99.57
Không RDRSegmenter	74.45	92.83	95.61	98.72	99.22

→ Bộ tách từ triệt tiêu **sự nhấp nhằng trong khoảng trắng** của tiếng Việt, tăng hiệu quả huấn luyện cho bài toán truy vấn.

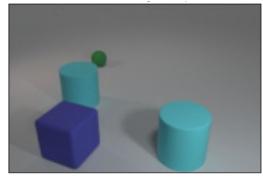
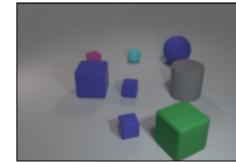
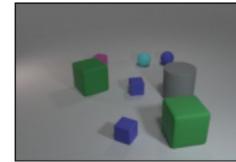
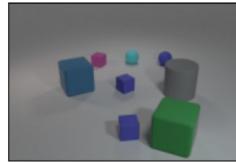
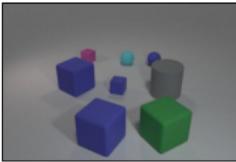
Kết luận

- 1 Áp dụng phương pháp **TIRG** trên **dữ liệu tiếng Việt** cho kết quả tốt hơn tiếng Anh do lợi thế đặc thù của tiếng Việt.
- 2 Sử dụng **PhoBERT** làm **Word Embedding** cho phép TIRG **thích ứng** với những từ **nằm ngoài dữ liệu huấn luyện**
- 3 Bộ tách từ **RDRSegmenter** loại bỏ sự **nhập nhằng khoảng trắng** ở tiếng Việt, tăng hiệu quả huấn luyện

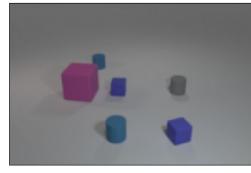
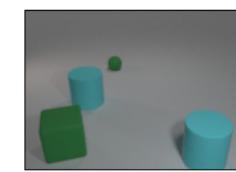
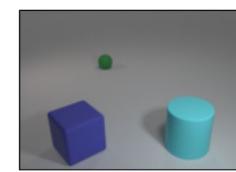
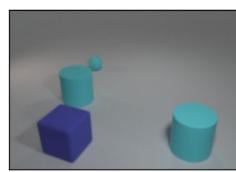
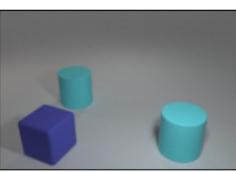
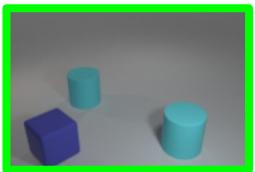
Kết quả truy vấn mẫu



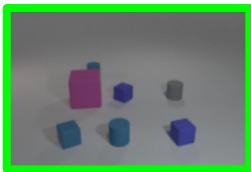
"biến khối lập phương nhỏ thành lớn"



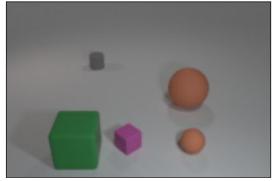
"xóa khối màu xanh lá góc trái phía trên"



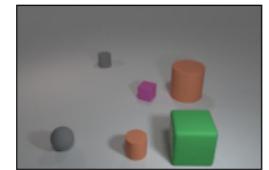
"thêm khối nhỏ vào góc trái phía dưới"



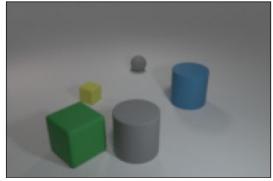
Kết quả truy vấn mẫu



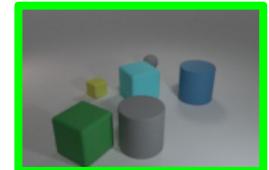
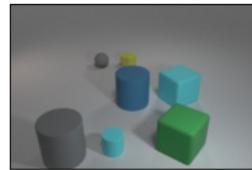
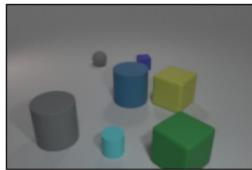
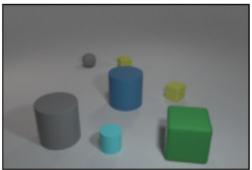
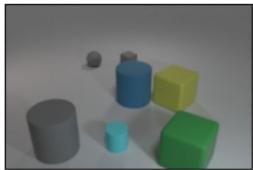
"thêm khối nhỏ
vào trung tâm"



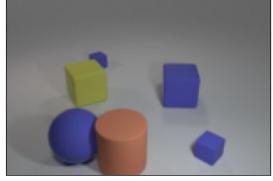
→ Hạn chế của mô hình biểu diễn ảnh



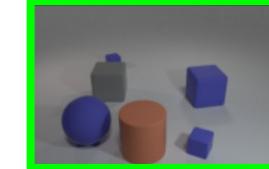
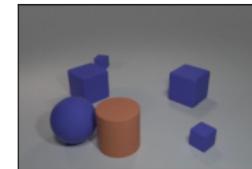
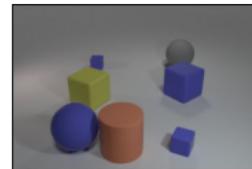
"thêm khối lấp
phương màu lam"



→ Hạn chế của hàm kết hợp



"biến khối vàng lớn
thành màu xám"

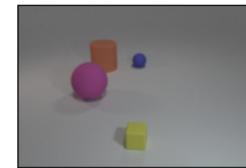
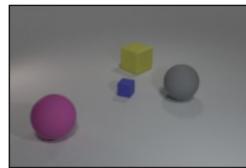


→ Hạn chế của mô hình biểu diễn ảnh

Kết quả truy vấn mẫu



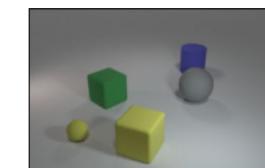
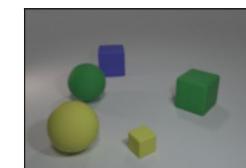
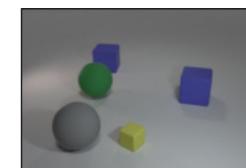
“xóa khối lập phương bên phải”



“xóa khối nhỏ”

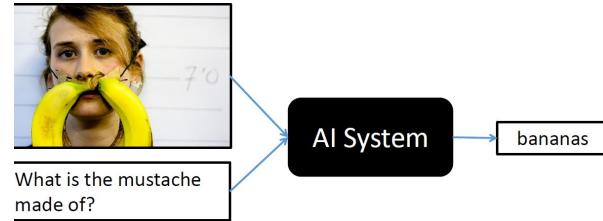


“biến khối màu vàng lớn thành màu đỏ”



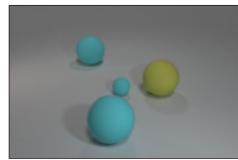
Hướng phát triển

- Phát triển kiến trúc mô hình
 - Bộ biểu diễn ảnh (EfficientNet,...)
 - Cải tiến hàm kết hợp $f_{combine}(\varphi_x, \varphi_t)$
- Mở rộng lên các bộ dữ liệu khác
 - MIT-States và Fashion200k
- Phát triển lên thành bài tập chí

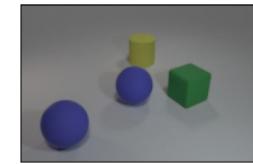
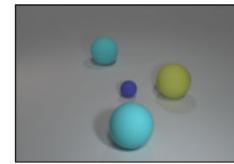
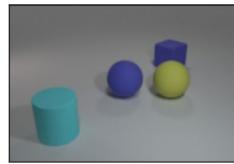
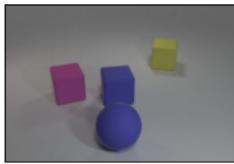
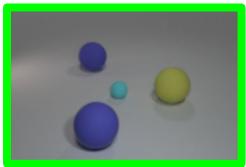




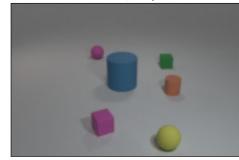
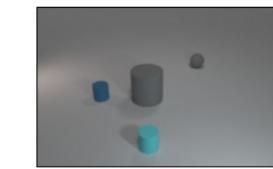
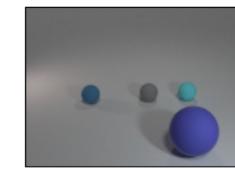
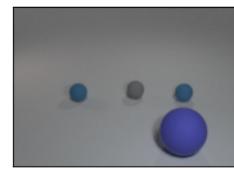
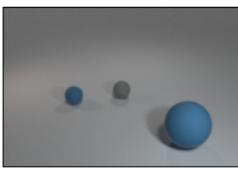
PHỤ LỤC



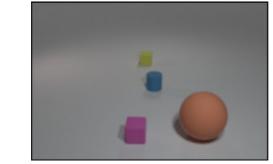
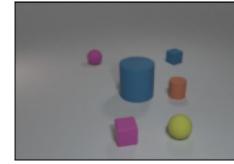
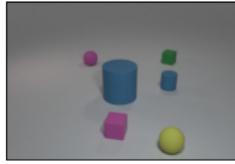
“Biến khối màu lam lớn thành
màu tím”

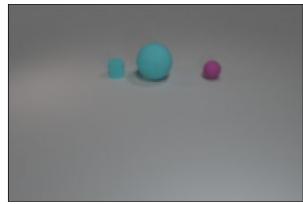


“Thêm khối trụ”

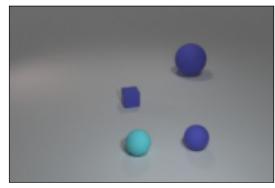
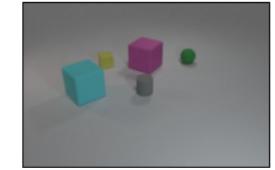
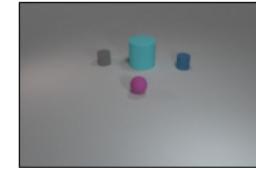
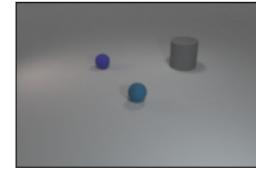
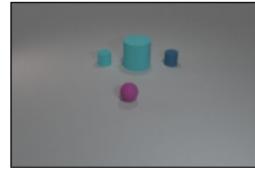


“Xóa khối trụ bên phải”

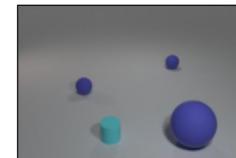
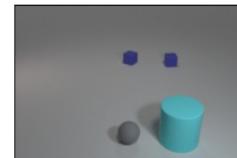
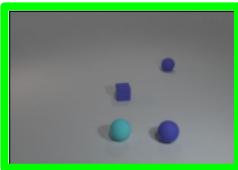




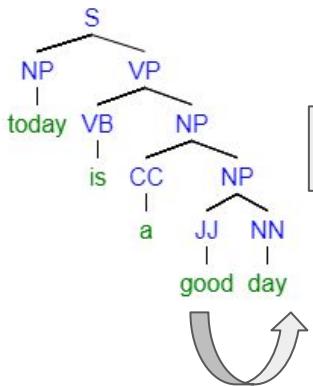
"thêm khối cầu
màu xám vào
trung tâm"



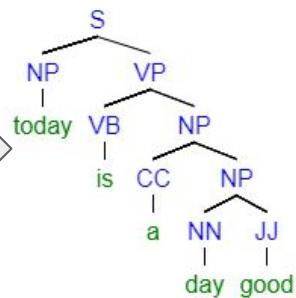
"biến khối màu
tím lớn thành
nhỏ"



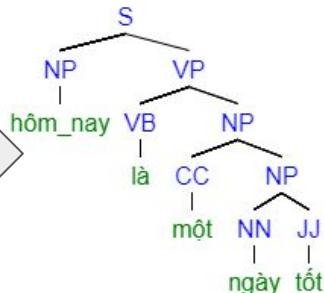
KẾT HỢP ẢNH VÀ CÂU MÔ TẢ TĂNG CƯỜNG TIẾNG VIỆT CHO TRUY VẤN ẢNH



Biến đổi về mặt cú pháp



Biến đổi về mặt từ vựng



Với mỗi batch b , chúng tôi tạo một tập \mathcal{N}_i bao gồm K mẫu: một mẫu ϕ_i^+ và $K - 1$ mẫu negative $\phi_1^-, \dots, \phi_{K-1}^-$ (bằng cách lấy mẫu ϕ_j từ minibatch với $j \neq i$.)

$$L = \frac{1}{MB} \sum_b^B \sum_m^M \log \left\{ \frac{\exp \{ \mathcal{K}(\psi_i, \phi_i^+) \}}{\sum_{\phi_j \in \mathcal{N}_b^m} \exp \{ \mathcal{K}(\psi_i, \phi_j) \}} \right\} \quad (26)$$

Trong đó \mathcal{K} là **hàm tương đồng** (*similarity kernel/function*) lấy vào 2 vecto.

Trong phương pháp chúng tôi sử dụng thì \mathcal{K} sử dụng **tích vô hướng** và **hàm đối l2** (*negative l2*).

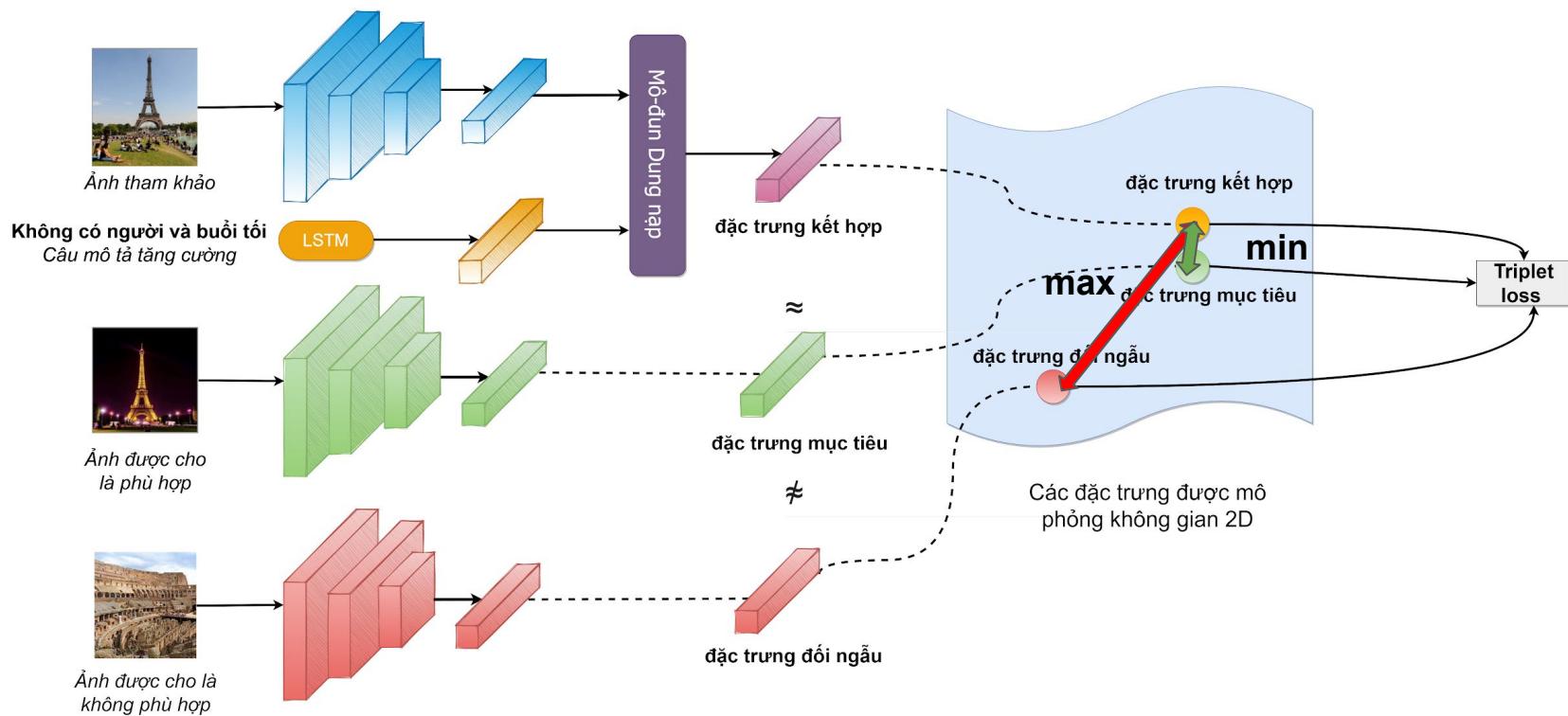
Khi $K = 2$ chúng ta có thể viết lại hàm L trên thành:

$$L = \frac{1}{MB} \sum_i^B \sum_m^M \log \{ \exp \{ \mathcal{K}(\psi_i, \phi_i^+) \} - \exp \{ \mathcal{K}(\psi_i, \phi_j^-) \} \} \quad (27)$$

A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification, 2017.

N. N. Vo and J. Hays. Localizing and orienting street views using overhead imagery. In ECCV, 2016.

Huấn luyện mạng TIRG



Tại sao lại là tiếng Việt?

1. Bài toán có **tiềm năng sử dụng cao**, tuy nhiên chưa được nghiên cứu trên tiếng Việt.

Tại sao lại là tiếng Việt?

2. Tiếng Việt là ngôn ngữ **không biến tố** (non-inflection), là lợi thế cho mô hình Học máy

VD1:

Tiếng Việt không chia động từ như tiếng Anh. Ở tiếng Anh, “He” đi với “works” còn “They” đi với “work”

VD2:

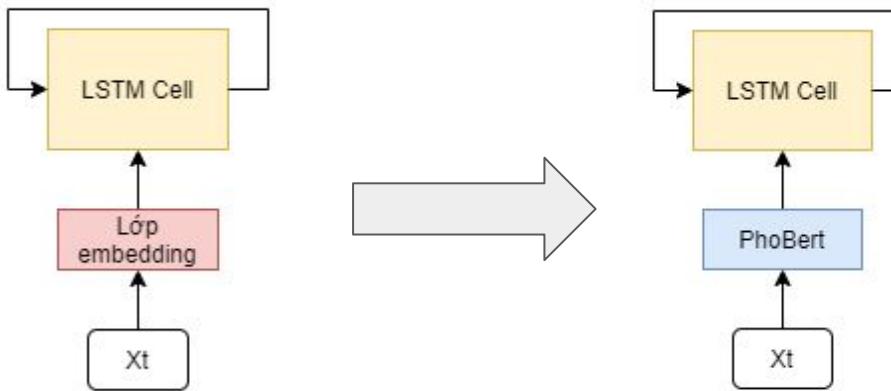
Tiếng Việt thì sử dụng “đã” để diễn tả hành động trong quá khứ thay vì biến tố “-ed”

→ Tiềm năng ứng dụng bài toán trên ngôn ngữ tiếng Việt

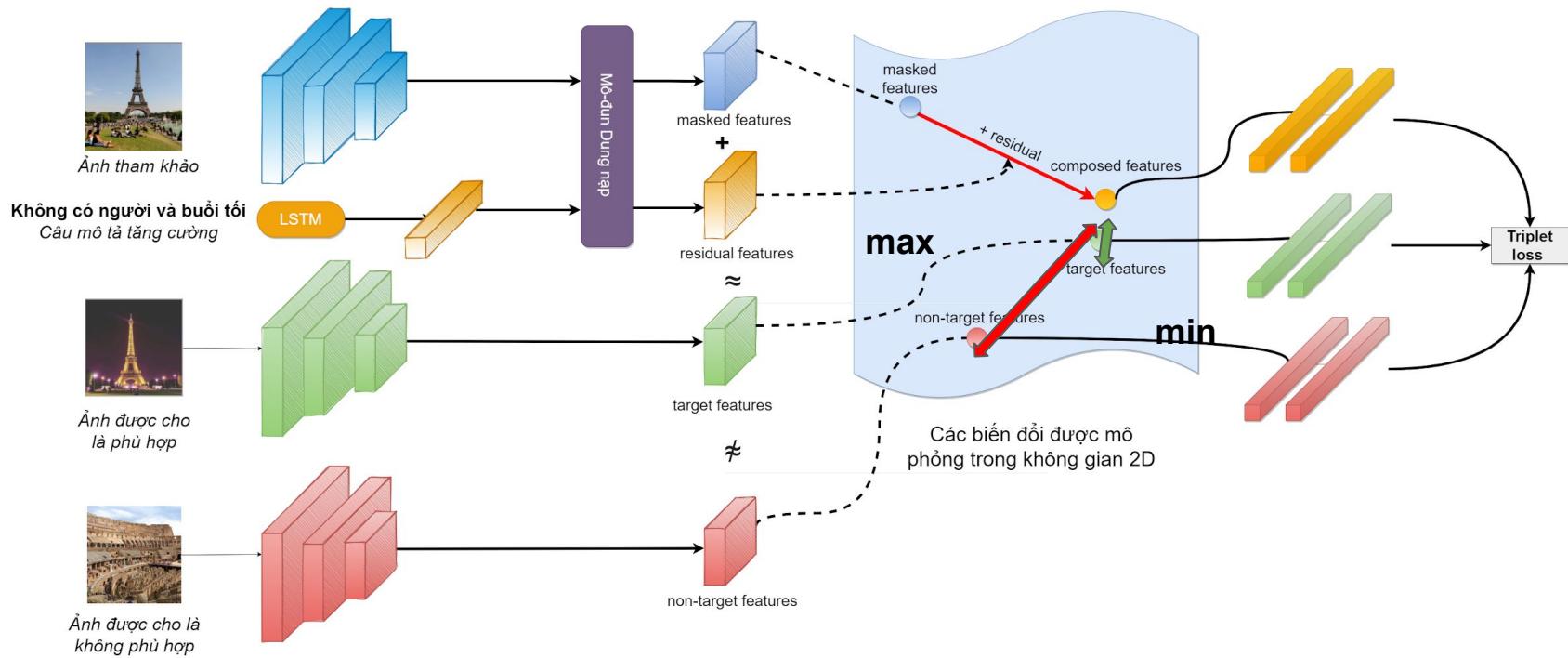
Đinh-Hòa Nguyễn, "Tiếng Việt không son phấn", John Benjamins Publishing Company, 1997.

Cải tiến TIRG để thích ứng với câu mô tả mới

Sử dụng **PhoBERT** làm **Word Embedding** sẽ **tăng độ thích ứng** của mô hình với câu mô tả chứa từ **năm ngoài dữ liệu huấn luyện**



TIRG



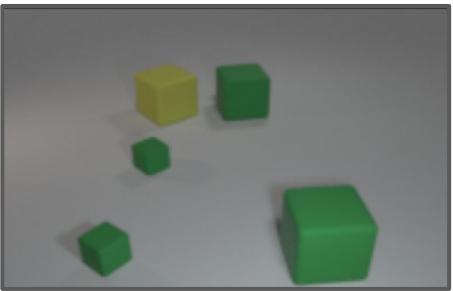
Nhận xét

Chúng ta có **rất nhiều phương pháp** để kết hợp biểu diễn ảnh và văn bản, tuy nhiên:

- Các phương pháp này được dùng để giải quyết các bài toán khác **không phải truy vấn ảnh**
- Các phương pháp này chỉ là những phép biến đổi cơ bản, do đó **không gian biểu diễn bị hạn chế**.
- Các phương pháp này xây dựng phép biến đổi bằng **chỉ sử dụng đặc trưng văn bản**, bỏ vai trò của **biểu diễn ảnh**.

Đặc điểm dataset

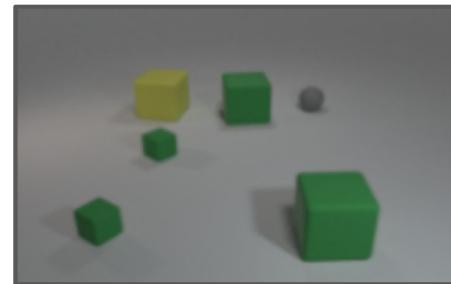
CSS



Biến đổi cục bộ

“Thêm khối màu xám”

TIRG-Conv



MIT-States



Biến đổi toàn cục

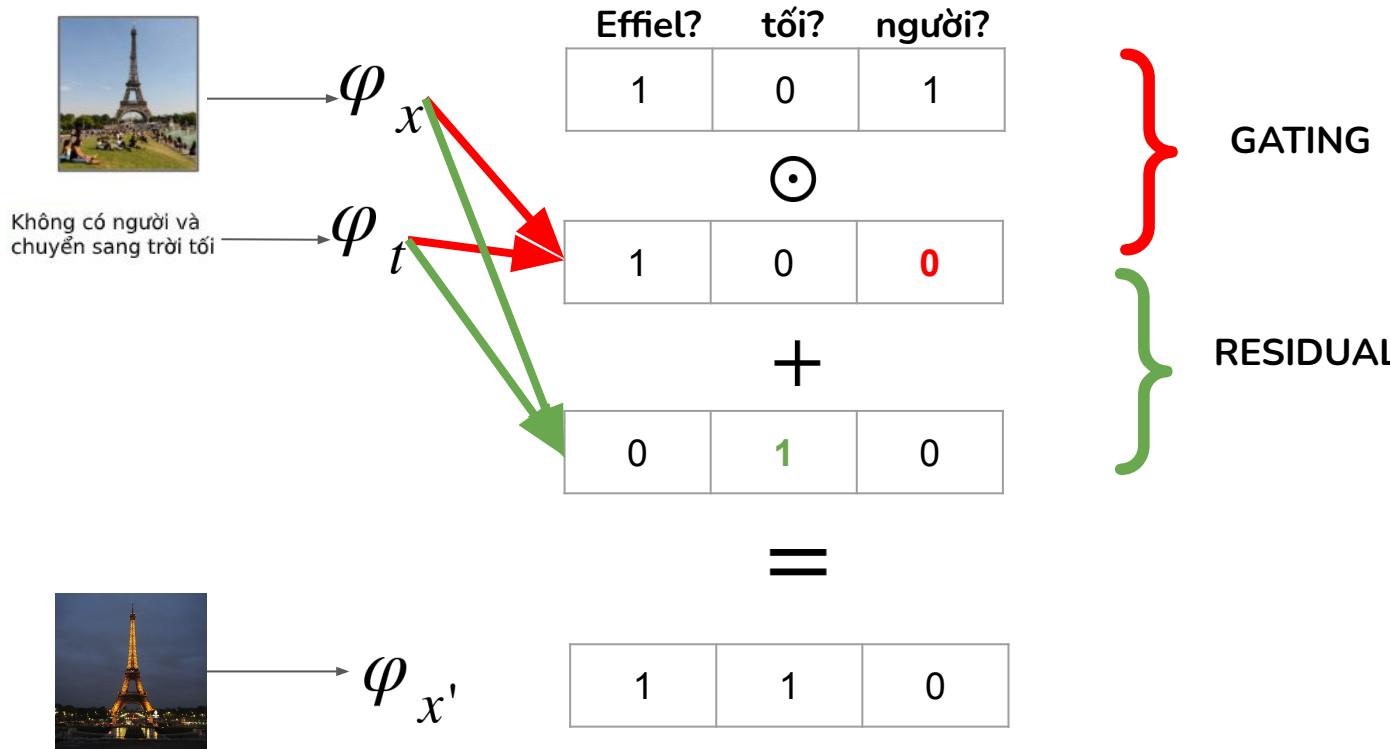
“Cũ kĩ”

TIRG-FC



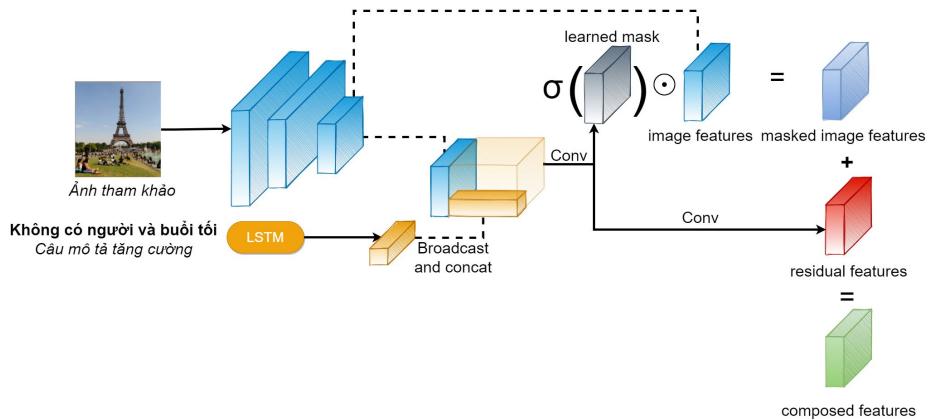
1. Giới thiệu bài toán

Ý tưởng chính Text Image Residual Gating

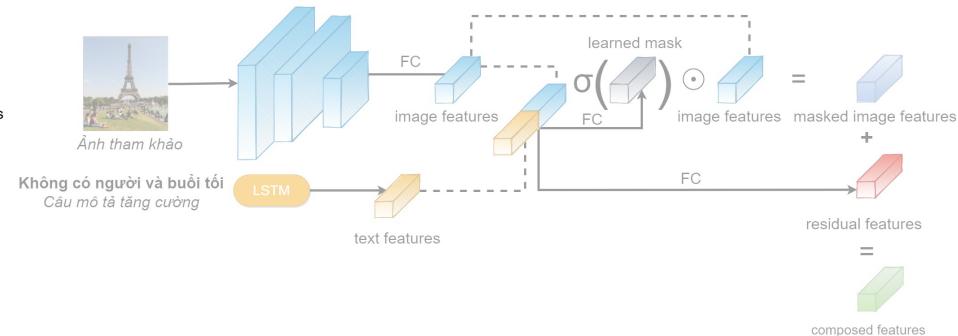


Hai mô hình của Text-Image Residual Gating

TIRG-Conv



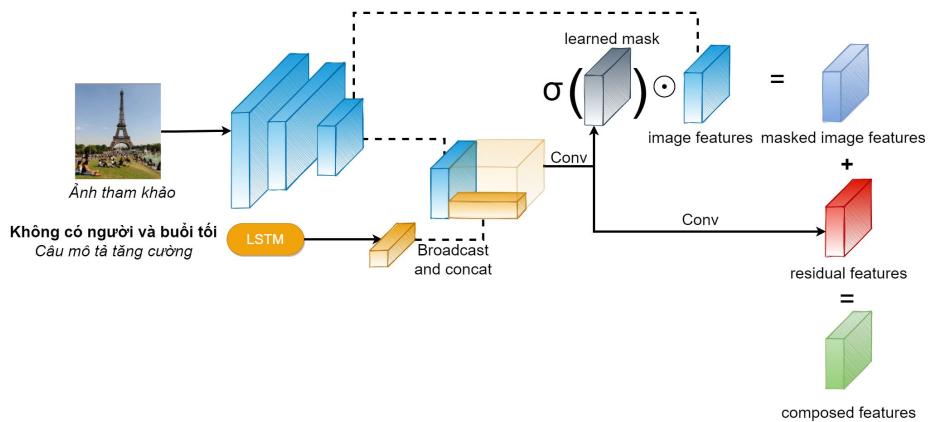
TIRG-FC



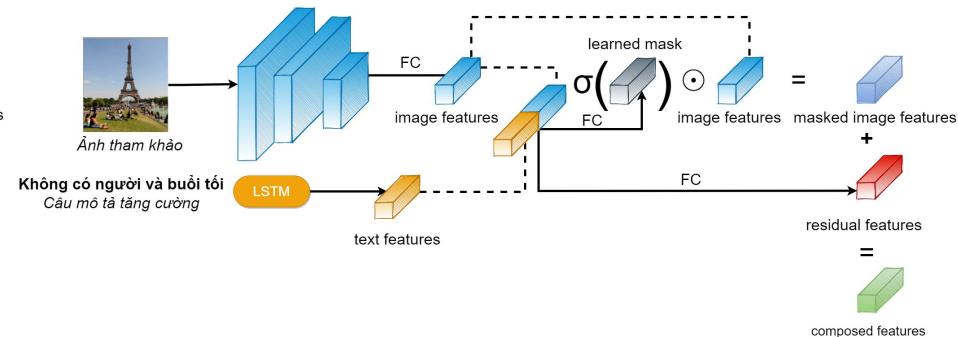
Vo, Nam, et al. "Composing text and image for image retrieval—an empirical odyssey." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

Hai mô hình của Text-Image Residual Gating

TIRG-Conv



TIRG-FC



Vo, Nam, et al. "Composing text and image for image retrieval-an empirical odyssey." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

Thí nghiệm 1. Kết quả huấn luyện mô hình TIRG trên tập CSS-VN

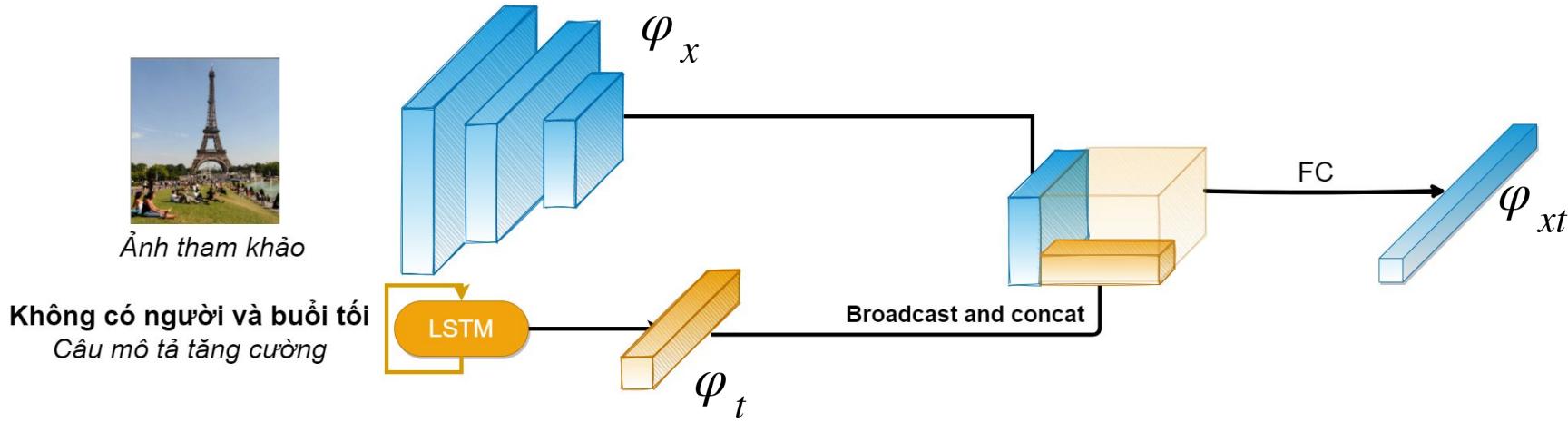
Cấu hình	R@1	R@5	R@10	R@50	R@100
TIRG-FC	78.03	94.91	97.18	98.28	99.57
TIRG-Conv	75.29	92.6	95.48	98.69	99.29

→ **TIRG-FC** cho kết quả **tốt hơn** **TIRG-Conv** trên tập **CSS-VN**, do đó chúng tôi sử dụng mô hình **TIRG-FC** để cải tiến

2. Các hướng nghiên cứu liên quan

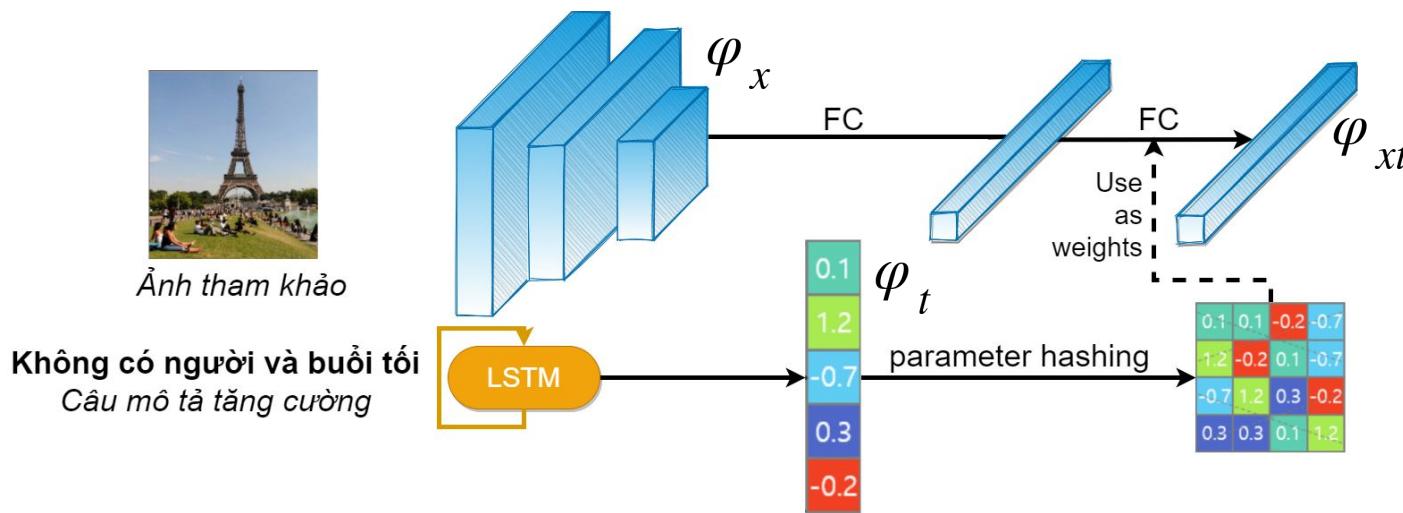
3. Hướng tiếp cận

Concatenation



1. S. Antol, et. al. VQA: Visual Question Answering, 2015.
2. I. Misra, A. Gupta, and M. Hebert. From red wine to red tomato: Composition with context.
3. B. Zhao, et. al. Memory-augmented attribute manipulation networks for interactive fashion search, 2017.

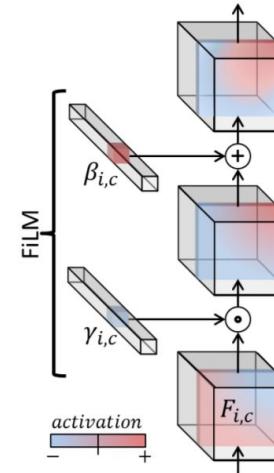
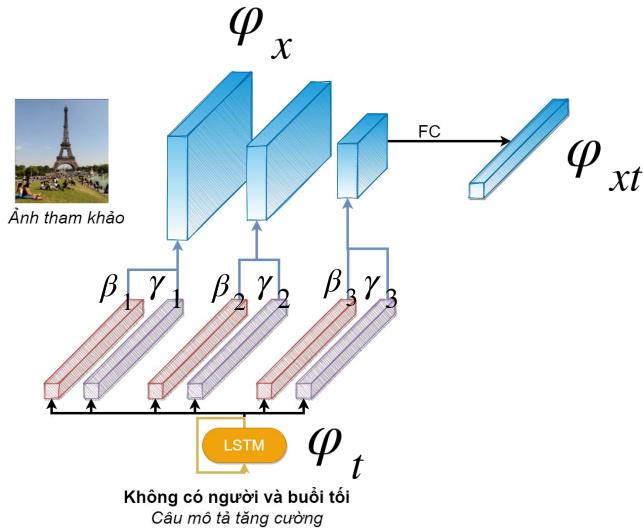
Parameter Hashing



- 1.. Phép biến đổi **đơn giản**, sử dụng ít tham số, do đó **không gian biểu diễn hạn chế**.
2. Việc sinh ra ma trận biến đổi chỉ **dựa vào biểu diễn văn bản**, bỏ qua **vai trò** của **biểu diễn ảnh**

H. Noh, P. Hongseok Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. In CVPR, 2016.

Feature-wise Linear Modulation (FiLM)



1. Phép biến đổi **đơn giản**, theo bài báo, chỉ có thể thực hiện một số phép như **lấy ngưỡng** (*thresholding*), **tỉ lệ** (*scaling*) và **phủ định** (*negating*).
2. Việc sinh ra các vectơ biến đổi chỉ **dựa vào biểu diễn văn bản**, bỏ qua **vai trò** của **biểu diễn ảnh**

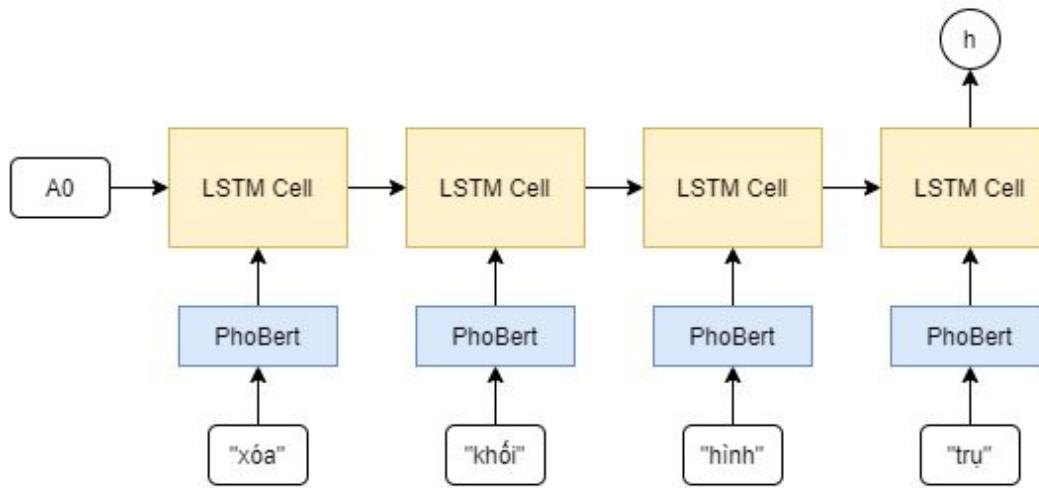
E. Perez, et. al. Film: Visual reasoning with a general conditioning layer. 2018.

5. Kết luận và hướng cải tiến

Thí nghiệm 2. Cải tiến mô hình bằng cách sử dụng PhoBERT bằng bộ biểu diễn từ

Thí nghiệm 1. n mô hình TIRG trên tập dữ liệu tiếng Việt CSS-VN

Sử dụng PhoBERT cho WordEmbedding



Ở mỗi thời điểm, các từ sẽ đi qua lớp Embedding là PhoBERT trước khi vào LSTM Cell

Thí nghiệm 2. Nghiên cứu cắt bỏ về các mô-đun kết hợp thay thế

Kết quả sử dụng các mô-đun kết hợp thay thế trên CSS-VN

	R@1	R@5	R@10	R@50	R@100
Chỉ dùng ảnh	06.59	29.04	53.08	94.09	96.51
Chỉ dùng câu mô tả	0.171	0.504	0.863	2.31	3.611
Concatenation	69.09	90.00	93.83	98.22	99.00
TIRG-FC	78.03	94.91	97.18	98.28	99.57
TIRG-Conv	75.29	92.6	95.48	98.69	99.29

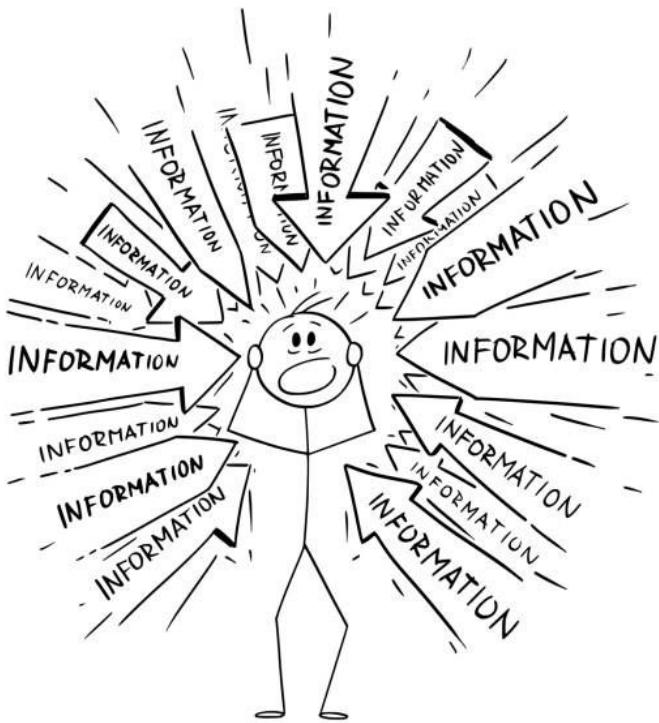
→ **TIRG** là một phương pháp **hiệu quả** để **biểu diễn cặp ảnh và câu mô tả tăng cường**.

Cấu hình huấn luyện

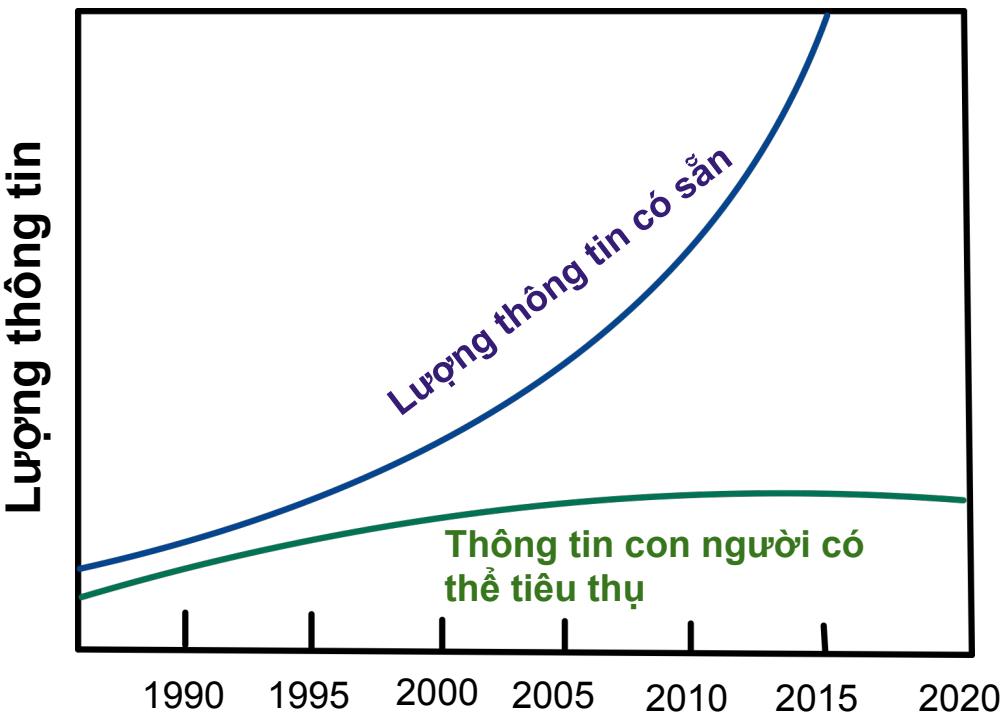
Cấu hình	Tham khảo	Sử dụng
Tốc độ học (LR)	0.01	0.01
Số lần lặp	160,000	594,000
Kích thước batch	32	32

→ **Huấn luyện dài hơi hơn** thời gian mà bài báo đề xuất, để mô hình có thể **khớp hơn với dữ liệu**

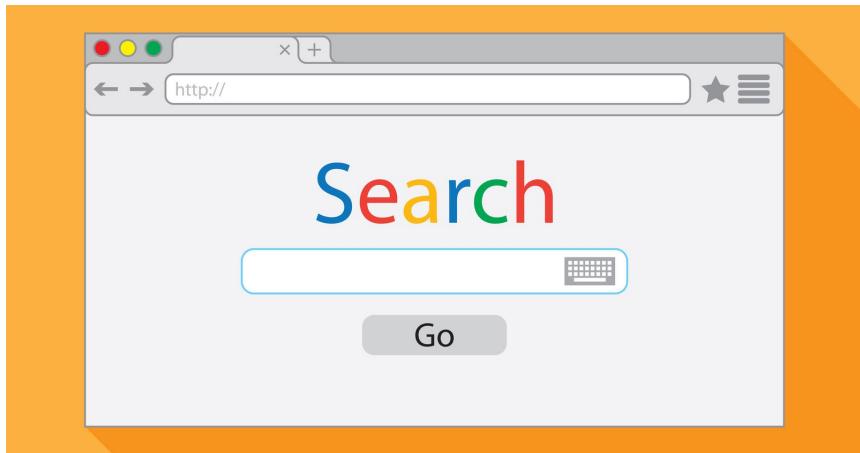
Sự phát triển quá nhanh của internet



Quá tải thông tin



Nhu cầu tìm kiếm ảnh



Google

thiên nhiên

All Images Maps Videos News More

hình nền tranh rừng background

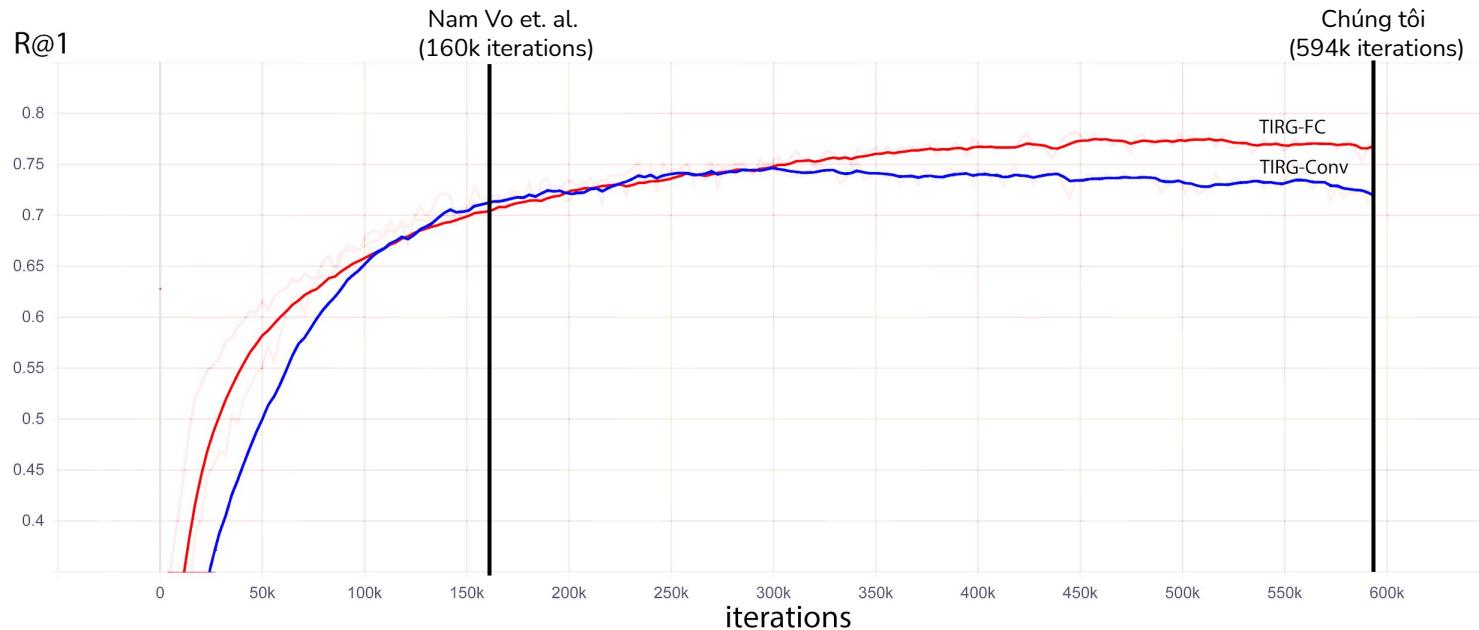
Mẹ Thiên Nhiên đã dạy chúng ta bài học gì?
phatgiao.org.vn

ThienNien.Net ...
thienhien.net

Một số kết quả truy vấn mẫu



So sánh TIRG-FC và TIRG-Conv



→ Khi huấn luyện dài hơi hơn thời lượng mà bài báo đề xuất, **TIRG-FC** cho kết quả **tốt hơn TIRG-Conv** trên **CSS-VN**

Thí nghiệm 1. Tái hiện lại mô hình TIRG trên tập dữ liệu tiếng Anh CSS

Kết quả tái hiện lại mô hình TIRG trên tập dữ liệu CSS

		Cấu hình	R@1	R@5	R@10	R@50	R@100
Bài báo công bố	TIRG-Conv	73.7	KCB*	KCB	KCB	KCB	
	TIRG-FC	71.2	KCB	KCB	KCB	KCB	
Kết quả tái hiện	TIRG-Conv	71.11	90.62	94.13	98.19	99.09	
	TIRG-FC	70.77	91.14	94.54	98.43	99.17	

*KCB: Không công bố

Truy vấn ảnh dựa trên đối thoại

Cải tiến TIRG để phục vụ cho bài toán truy vấn ảnh dựa trên đối thoại. Khi đó φ_{xt} sẽ liên tục được cập nhật với biểu diễn của câu mô tả mới của người dùng tại thời điểm t : φ_t để cung cấp hệ thống về nhu cầu thông tin



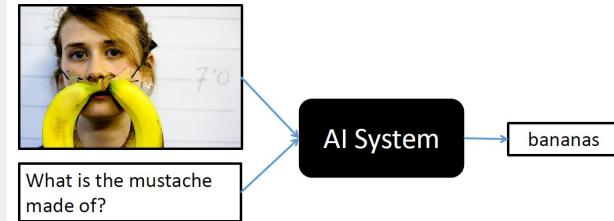
Xiaoxiao Guo*, Hui Wu*, Yu Cheng, Steven Rennie, Gerald Tesauro, Rogerio Schmidt Feris.

“Dialog-based Interactive Image Retrieval.” NeurIPS 2018

Sử dụng TIRG cho các bài toán khác

Cải tiến **TIRG** để phục vụ cho các bài toán khác cần kết hợp biểu diễn ảnh và văn bản như:

- **Hỏi đáp dựa trên ảnh (Visual Question Answering)**
- **Chỉnh sửa ảnh dựa trên đối thoại (Conversational Image Editing)**



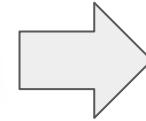
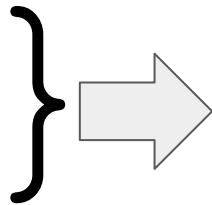
(i) S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In ICCV, 2015.

(ii) Manuvinakurike, Kallirroi. "Conversational Image Editing: Incremental Intent Identification in a New Dialogue Task.", 2018.

Ngữ cảnh ứng dụng



Ngữ cảnh ứng dụng



“Giống vậy nhưng nó màu xanh”